

AI and cybersecurity: a delicate balance

Data vs Knowledge

Maria Vanina Martinez
ISODARCO 2025, Volterra



Cybersecurity



Cybersecurity – *General Definition*

Protecting Internet-connected systems from attacks, including software, hardware, data, and people.

This includes many aspects:

- Application security
- Information security (including things like fake news)
- Network security (includes things like malicious actors)
- Disaster recovery
- Operational safety
- User education (phishing, etc.)



Cybersecurity - Motivation

- Dynamism
- Heterogeneity
- Uncertainty
- Incompleteness
- Decision support (socio-technical) systems can benefit from the use of intelligent software systems.



Where can AI come in?

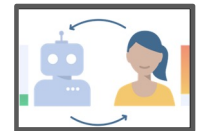


Opportunities

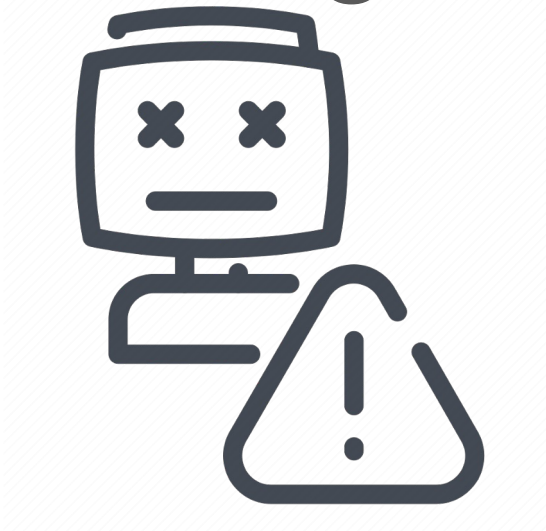
AI plays an important role in many cybersecurity challenges:

- Cybercrime prevention
- Defense of corporate networks
- Automatic generation of alerts based on online content
- *patching*
- Identification of malicious actors (bots, fake news, trolls, etc.)
- Threat hunting using human-in-the-loop systems

...and much (much!) more.



What could go wrong?



What could go wrong...?

Two major sources of threats:

- *Malicious use* of Artificial Intelligence.
- *Vulnerabilities* introduced by the expansion of the attack surface when incorporating AI into information systems.

Some malicious uses of AI



Threats

Spam

Fake news

Social engineering

Unauthorized identification
(facial recognition, etc.)

Creation of fake content
(audio, image, video, websites, ...)

Phishing

Bots and Botnets
(many possible uses...)

Malware Development

Automated hacking

Vulnerable intelligent systems:

Some examples



Threats

Poisoning
of data and models

Evasion Attacks

Entrance attacks

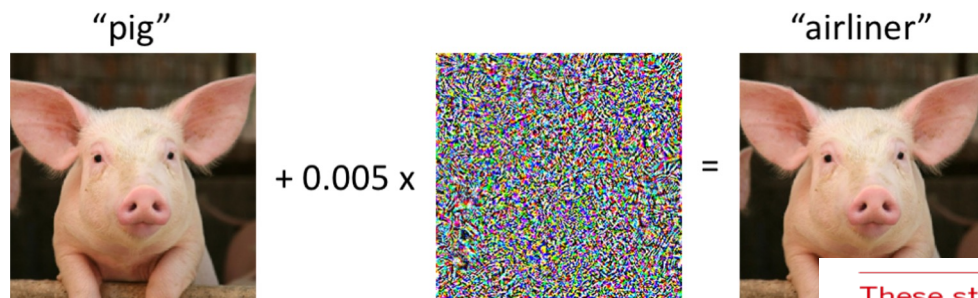
Training Data Theft

Prompts Injection

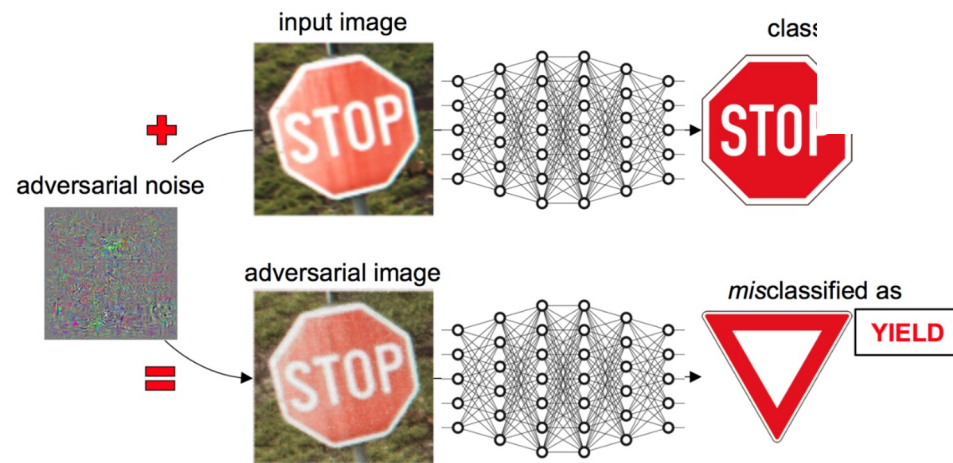
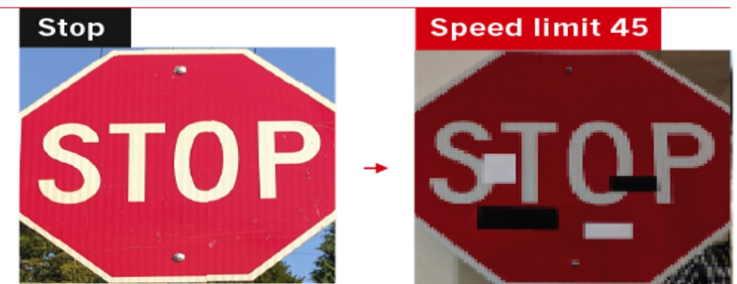
Model Theft

AI attacks are fundamentally different

- Unlike traditional cybersecurity vulnerabilities, the problems that create AI attacks cannot be “fixed” or “patched.”
- Traditional cybersecurity vulnerabilities are generally a result of programmer or user error. As a result, these errors can be identified and rectified.
- In contrast, the AI attack problem is more intrinsic: **the algorithms themselves and their reliance on data.**
- Furthermore, interconnection and autonomy also increases the surface and opportunities for attacks.



These stickers made an artificial-intelligence system read this stop sign as 'speed limit 45'.



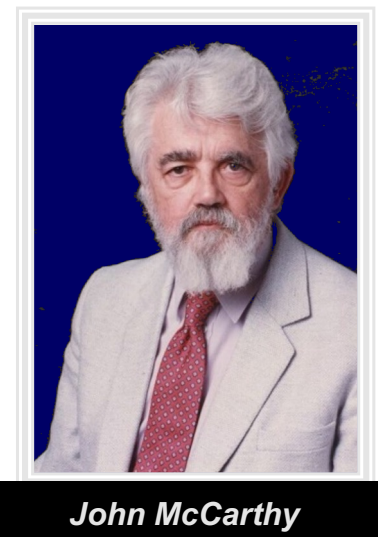
What is Artificial Intelligence?

- The AI we imagine(d)
- The AI we have
- The AI we need



What is Intelligence?

- It is the faculty or ability to *understand* and *reason* with facts, truths or propositions. Rationality. (Dictionary)
- the ability to apply knowledge to manipulate one's environment or to think abstractly as measured by objective criteria (such as tests).
- The first is associated to the idea “*learning*”, the second to the idea of “*problem solving*” or “*goal oriented behavior*”.
- Intelligence is the *computational part* of the ability to achieve goals in the world. (John McCarthy)



Some attributes of our model of intelligence

- Ability to *learn*.
- Ability to *solve problems* (break down complex problems into simpler ones).
- Ability to *make plans and predict the consequences of the actions* considered (ability to evaluate and compare alternatives).
- Originality: *synthesize new concepts* and ideas and be able to develop and use *analogies*.
- Generalize: find a *common underlying pattern* in a set of apparently different/similar situations.
- Perceive and *adequately model the world* (according to the problem that needs to be solved).
- *Understand and use language* and other symbolic tools.
- Having *mental attitudes*: beliefs, desires, intentions.
- *Ability of comprehension* (including making sense of ambiguous or contradictory information).
- *Understanding of the limits of own knowledge and skills*.

Different levels of Intelligence?

- Basic Cognitive processes:
 - Pattern recognition: faces, voices, words, etc.
 - Ability to speak, read, remember, etc...
- Complex Cognitive Processes:
 - Involve a combination of other processes
 - Includes conceptual learning, building abstractions, planning, argumentation, decision making, reasoning, etc...

Thinking Fast

“Fast thinking includes both variants of intuitive thought—the expert and the heuristic—as well as the entirely automatic mental activities of perception and memory, the operations that enable you to know there is a lamp on your desk or retrieve the name of the capital of Russia.” [1] Thinking Fast and Slow – D. Kahneman

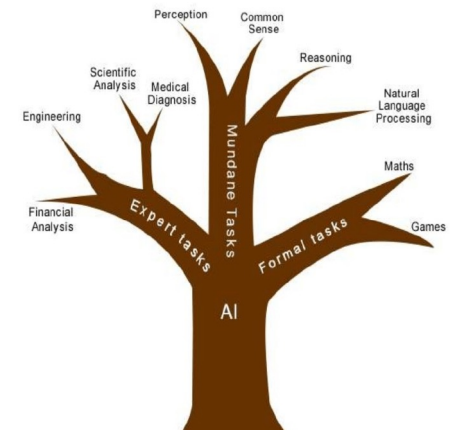
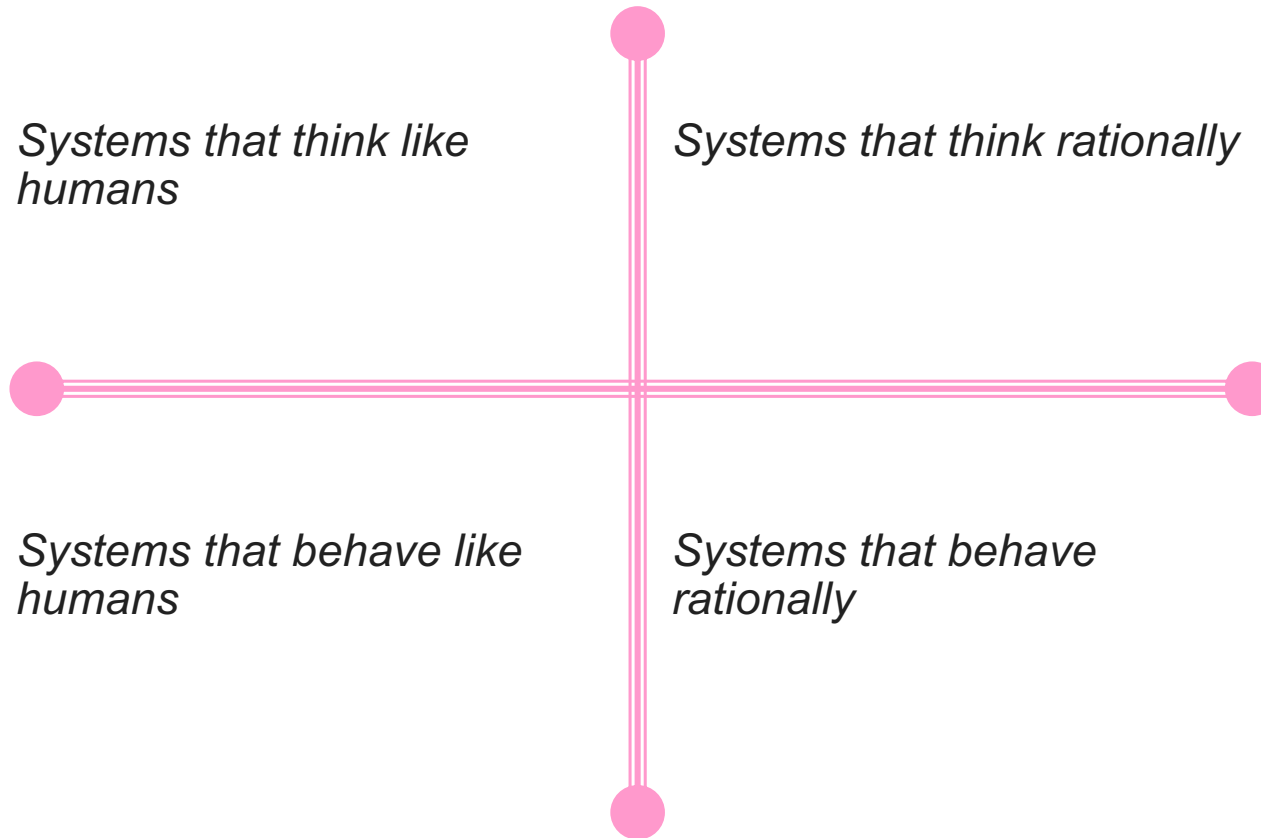
- Intuitive thought: e.g., the recognition of a face
- Recognize specific characteristics: facial characteristic, color of hair,
- Mood, e.g., this person looks angry/upset
- Predict the future: eg., this person is about to scream



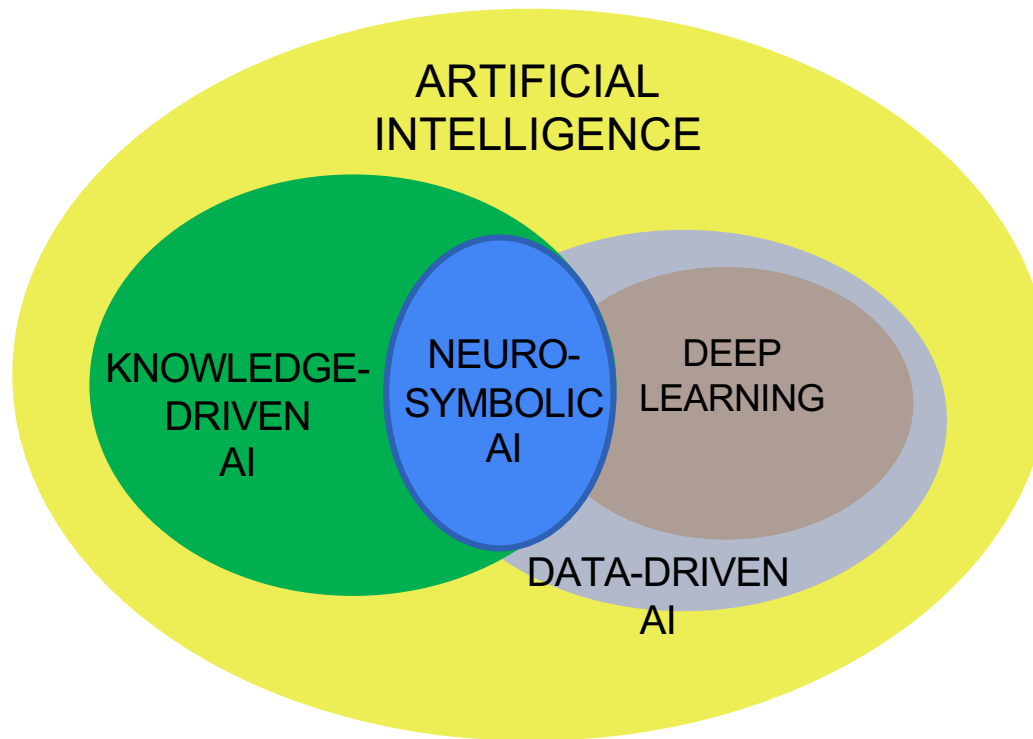
Thinking Slow

- “The spontaneous search for an intuitive solution sometimes fails— neither an expert solution nor a heuristic answer comes to mind. In such cases we often find ourselves switching to a slower, more deliberate and effortful form of thinking. This is [...] slow thinking [...]” [1]
- $23 * 14 = ?$
- Intuitive thought: quickly we recognize that there is an equation, it is a multiplication, a results is expected. We also know immediately that we could solve it and may guess a range of possible values for the result.
- Now.... To actually solve it, there is a series of “mental steps” (and sometimes even physical): we have to “retrieve” the procedure we once learned of how to multiply and apply it. It requires an *effort*, it is not immediate.

The AI we envisioned



The universe of AI



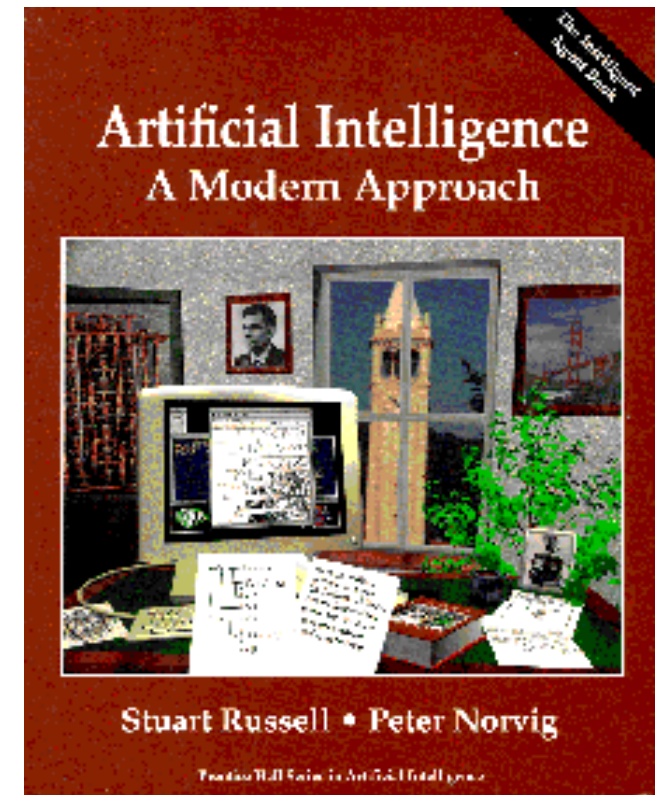
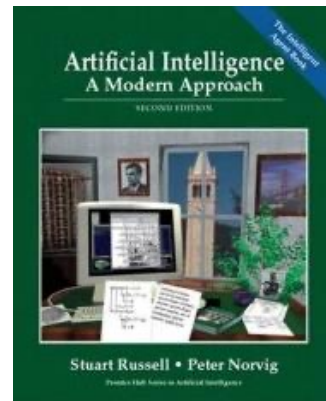
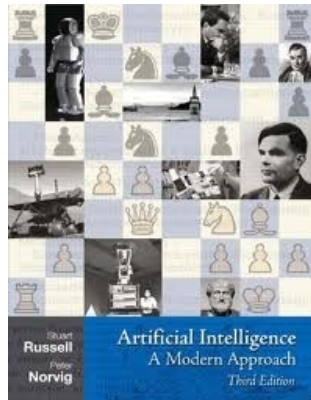
Reading Material

Artificial Intelligence: A Modern Approach

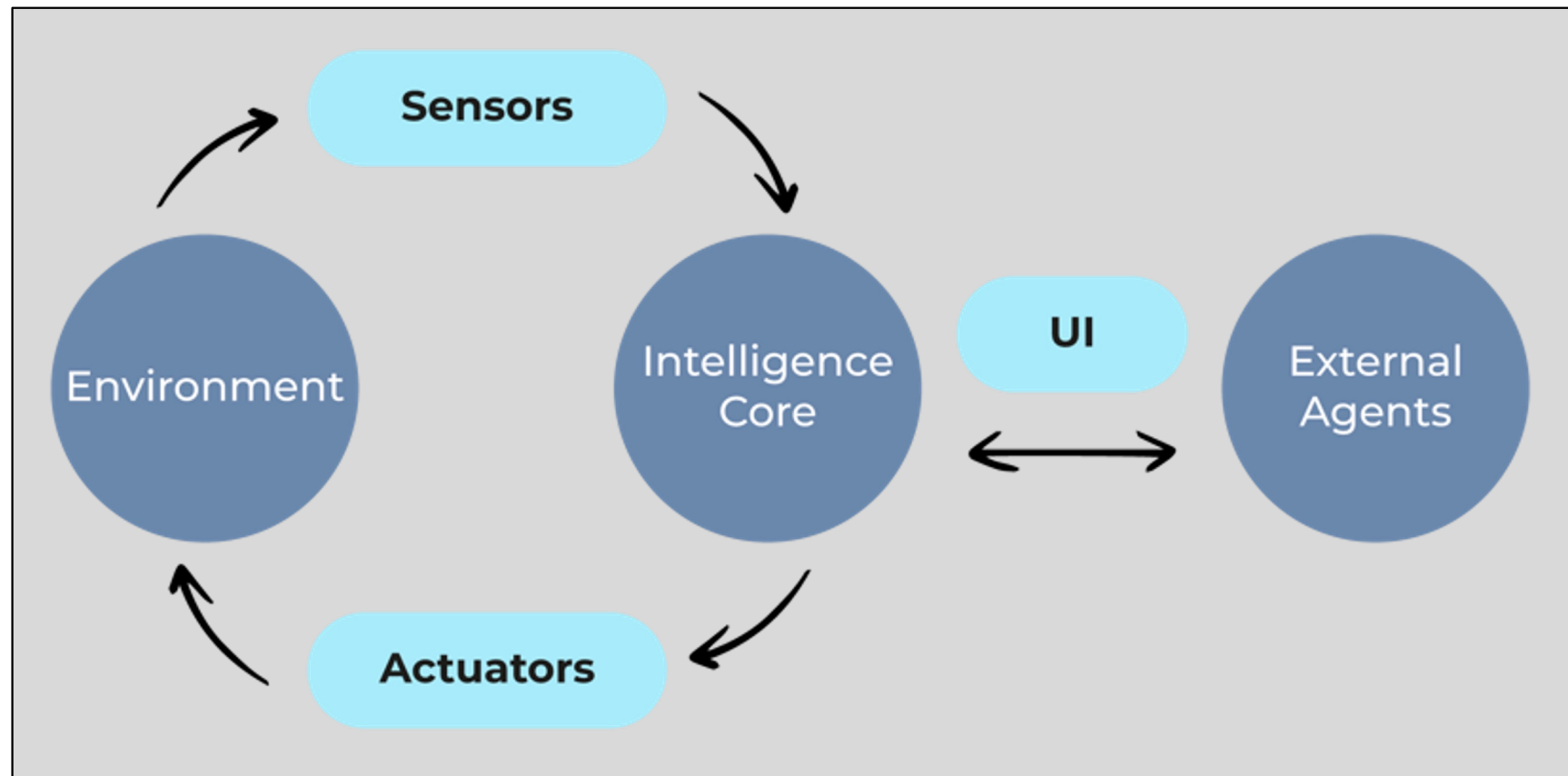
S. Russell, P. Norvig

Prentice Hall

4th Edition available.



From AI to *Intelligent Systems*



Data-driven AI

- Data is a record of things that exist or have happened and have an implicit meaning.
- For example, names, student ID numbers, phone numbers, employee numbers, student numbers, images, audio recordings, books, bank transactions, historical events, events in general, etc.
- We organize the data in databases as a collection of related data: excel, relational databases, the whole internet..
- A database is a logically coherent collection of data with a meaning that depends on the application domain and makes sense (only) to their users.



Data-driven AI



Seeks to emulate intelligent behavior that can be inferred from data (systems that act as the record we have of humans).



From data (more or less raw) statistically significant patterns are inferred. These correlations is what we call *information*.

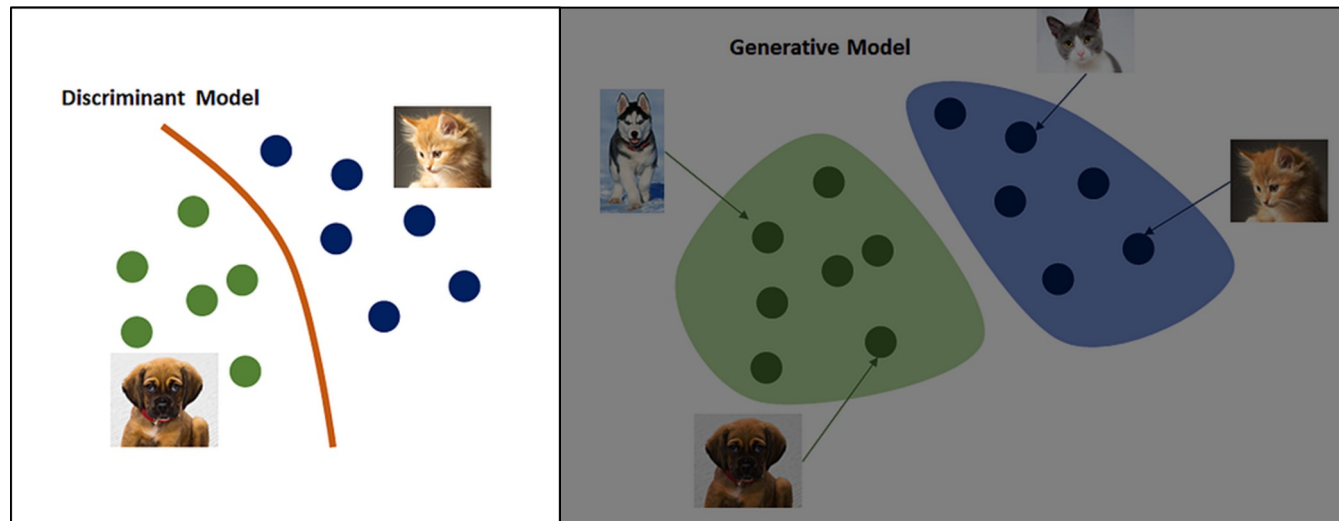


These patterns are used to define classification criteria, define processes for predicting phenomena, etc.



Examples: Neural networks, deep neural networks, association/decision trees, language models, etc.

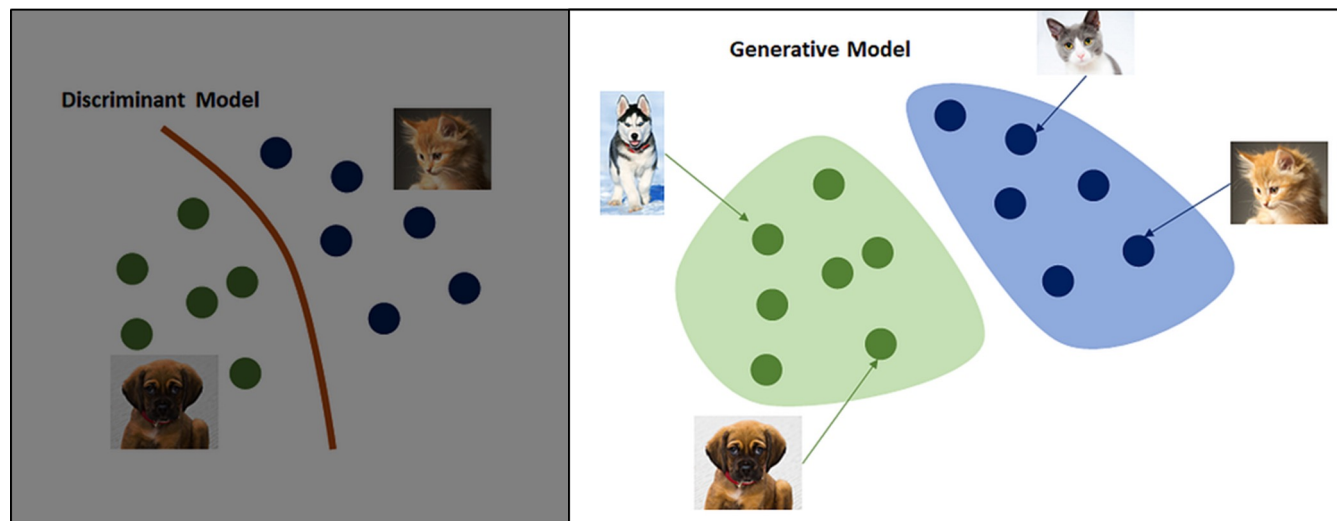
ML – Discriminative vs. Generative Models



Discriminative models (for classification tasks):

- They establish decision boundaries (find a function) to separate interest classes
- They do not need to describe the elements in the classes

ML – Discriminative vs. Generative Models



Generative models (for classification tasks):

- They construct probability distributions based on classes and characteristics of their elements.
- This gives us information about the content of each class.
- In particular, sampling from distributions “generates” elements.

Problems of data-driven AI

- Data is not always correct, does not always exist, it is difficult to curate for massive training, out-of-distribution data, distribution shift, unrepresentative, biased, etc.
- Reproduce patterns based on existing data, positive examples, learn from the past reflected in the data, find correlations, not causes.
- In general these models are neither interpretable, previsible, nor explainable.
- There is no place to represent explicit knowledge (e.g., what is true/false, common sense, etc).
- They do not generalize (retraining is often required).
- Unsupervised systems cannot be easily cured.
- These issues make AI systems particularly vulnerable.

Understanding in LLMs

- Potemkin understanding: the illusion of understanding driven by answers irreconcilable with how any human would interpret a concept.
- Study shows these failures reflect not just incorrect understanding, but deeper internal incoherence in concept representations.
- LMs do not truly understand language, beyond superficial inferences that are essentially the byproduct of the memorization of massive amounts of ingested text.
- Other studies identified logical inconsistencies through methods including: reversed claims, decision-making inconsistencies, factual correctness in paraphrases, compositional tasks, negation handling, common human misconceptions, semantic equivalence consistency benchmarks, and validation-generation alignment.

The image displays three chat interface examples, each with a user question and a GPT-4 response.

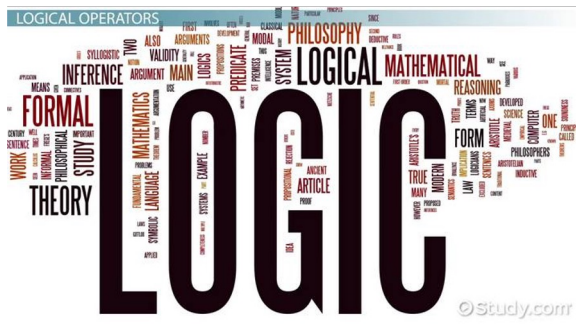
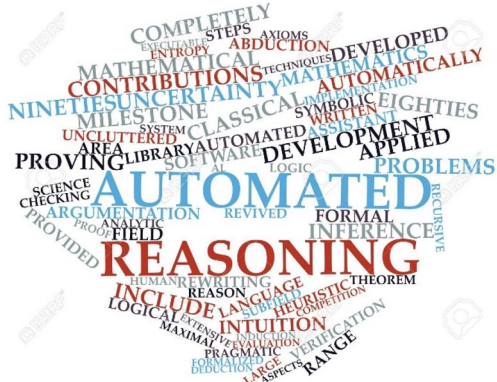
Example 1 (Correct): The user asks, "What is an ABAB rhyming scheme?". The model responds, "An ABAB scheme alternates rhymes: **first and third lines rhyme, second and fourth rhyme.**". The response is labeled "Correct".

Example 2 (Potemkin): The user asks, "Fill in the blank so that the resultant poem uses the ABAB rhyming scheme:". The poem provided is:
Wondrous winter calls **out**
Shivering under the frost
Lies a lonely cat, sitting ?
Alone but hardly lost
The model responds, "Lies a lonely cat, sitting **soft**". This response is labeled "Potemkin" in red, indicating a failure because "soft" does not rhyme with "out".

Example 3 (Correct): The user asks, "Does **out** rhyme with **soft**?". The model responds, "No". The response is labeled "Correct".

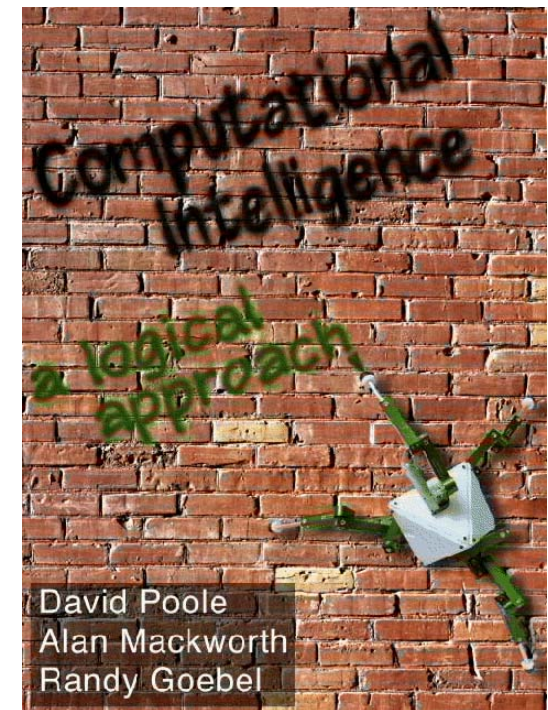
Mancoridis, M., Weeks, B., Vafa, K., & Mullainathan, S. Potemkin Understanding in Large Language Models. ICLM 2025.

Knowledge-driven AI



Computational Intelligence: A Logical Approach

D.Poole, A. Mackworth, R. Goebel
Oxford University Press.



Knowledge Representation

- We (humans) create models of parts of the world all the time so we can “reason” about them.
- A representation is an object that takes the place of another and highlights features that are relevant to the purpose of the representation.



Knowledge Representation and Reasoning

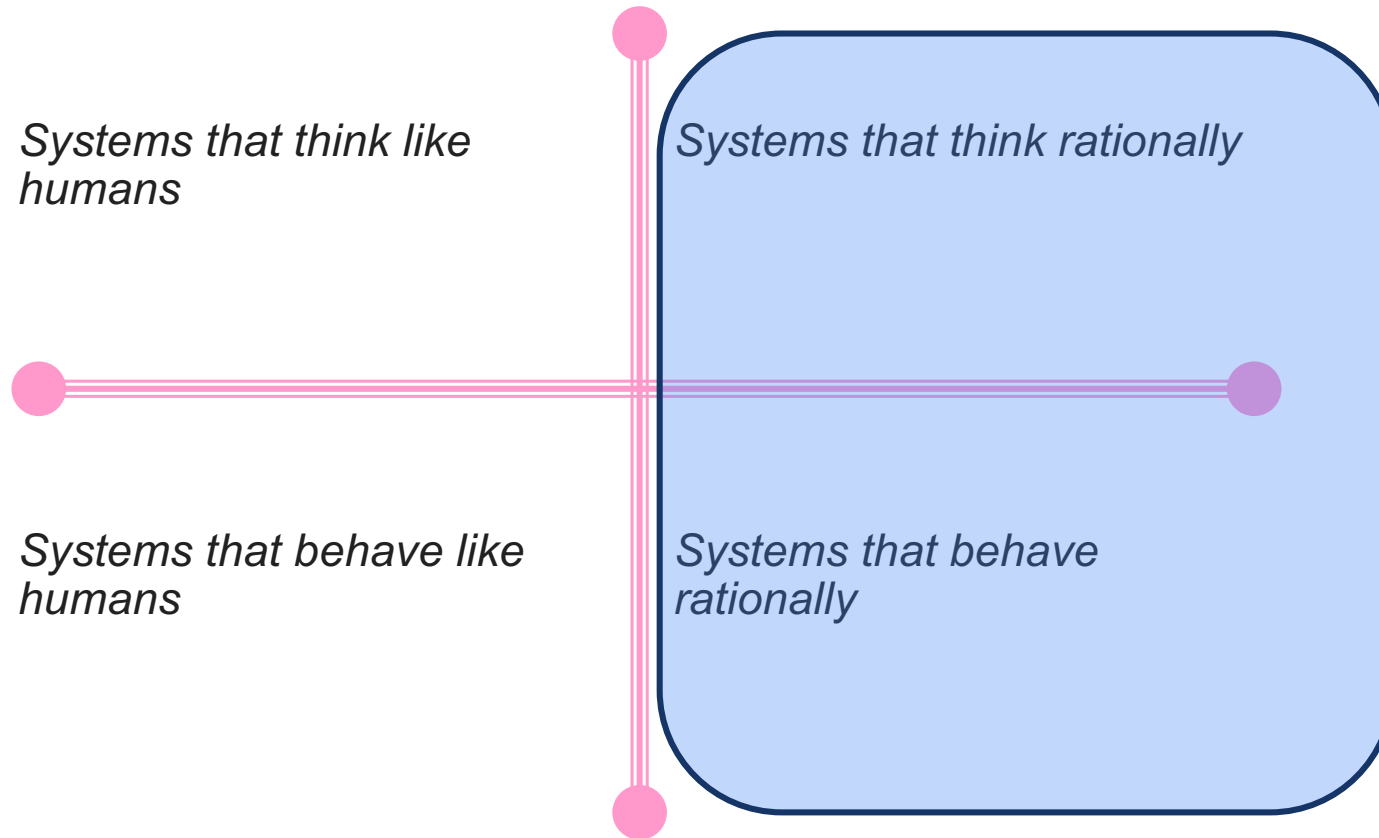
- Notion of intelligence: “human capacity to understand the world by creating symbolic representations.”
- Notion of knowledge: Knowledge is a skill or theoretical understanding of a topic/concept.

Assumption 1 : There is a set of rules (they can be very complex) that capture everyday human knowledge as manipulation (deduction, induction, abduction) of symbolic structures.

Assumption 2 : These rules can be discovered, formalized, and implemented on a computer (language and semantics beyond classical logical, capturing uncertainty, inconsistency, time reasoning).

Areas of research and development: KR&R, Automated Reasoning, Automatic Planning

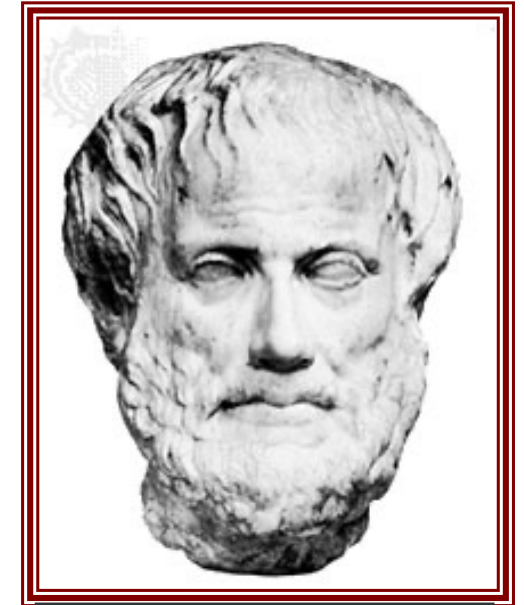
The AI we envisioned



Thinking rationally

Thought is considered to be governed by formal laws.

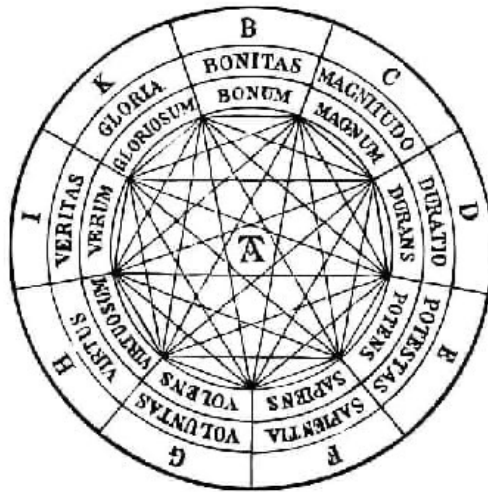
- Aristotle appears to have been the first to attempt to codify “the form of thought.”
- For him, “thinking” is a process and this process can be formalized.
- In fact, this way of seeing thinking has fueled both the progress of Formal Logic and that of Computer Science.



Aristoteles (384-322 ac)

Thinking rationally

- Ramon Llull (1232-1315): tried to use logic to explain the existence of the Holy Trinity.
- In doing so, he developed the first diagrammatic and mechanical tool for reasoning.



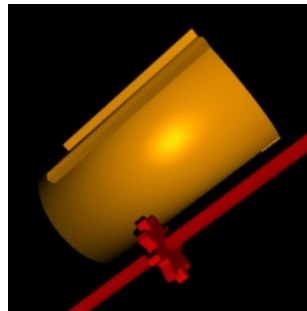
Source: https://en.wikipedia.org/wiki/Ramon_Llull



Ramon Lull
(1232-1315)

Thinking rationally

- Leibniz was the co-inventor of the integral calculus, also invented binary notation, mechanical calculator, etc.
- Inspired by Llull he dreamt of creating a universal language (*lingua characteristica universalis*) that would allow to express all human knowledge, and a set calculational rules (calculus ratiocinator), so a machine could derive all logical relationships.



https://en.wikipedia.org/wiki/Leibniz_wheel

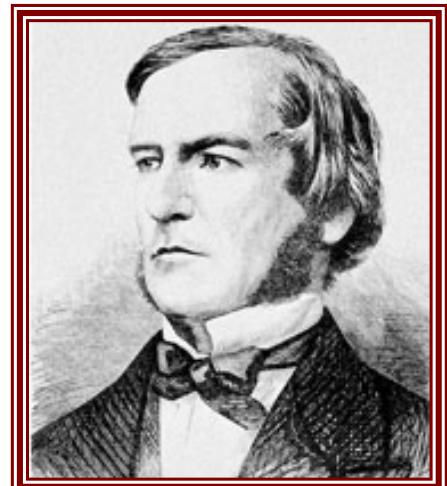


Gottfried Wilhelm Leibniz
(1646-1716)

Thinking rationally

In 1854, G.Boole published a seminal work in Mathematical Logic titled “An Investigation of the Laws of Thought on Which Are Founded the Mathematical Theories of Logic and Probabilities.”

- Studies logic from a mathematical point of view: algebraically
- Developed the Boolean Algebra
- This work gave an enormous boost to the formal treatment of Logic.
- Shannon (1937) connected Boolean Logic and circuits design.



George Boole
(1815-1864)

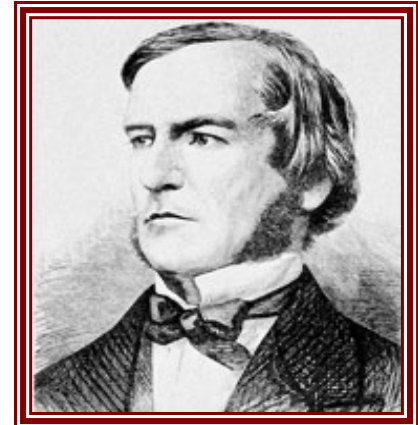
Thinking Rationally

World Logic Day Moshe Vardi
Lecture: What Came First,
Math or Computing?

<https://www.youtube.com/watch?v=CiD2Hm4GmpQ>



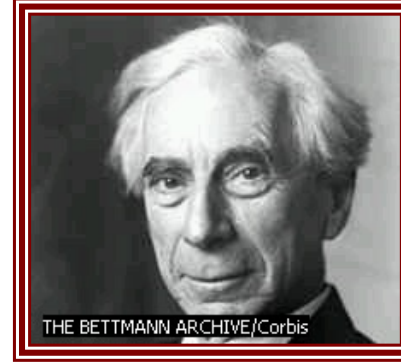
Gottfried Wilhelm Leibniz
(1646-1716)



George Boole
(1815-1864)



Hilbert, David
(1862-1943)

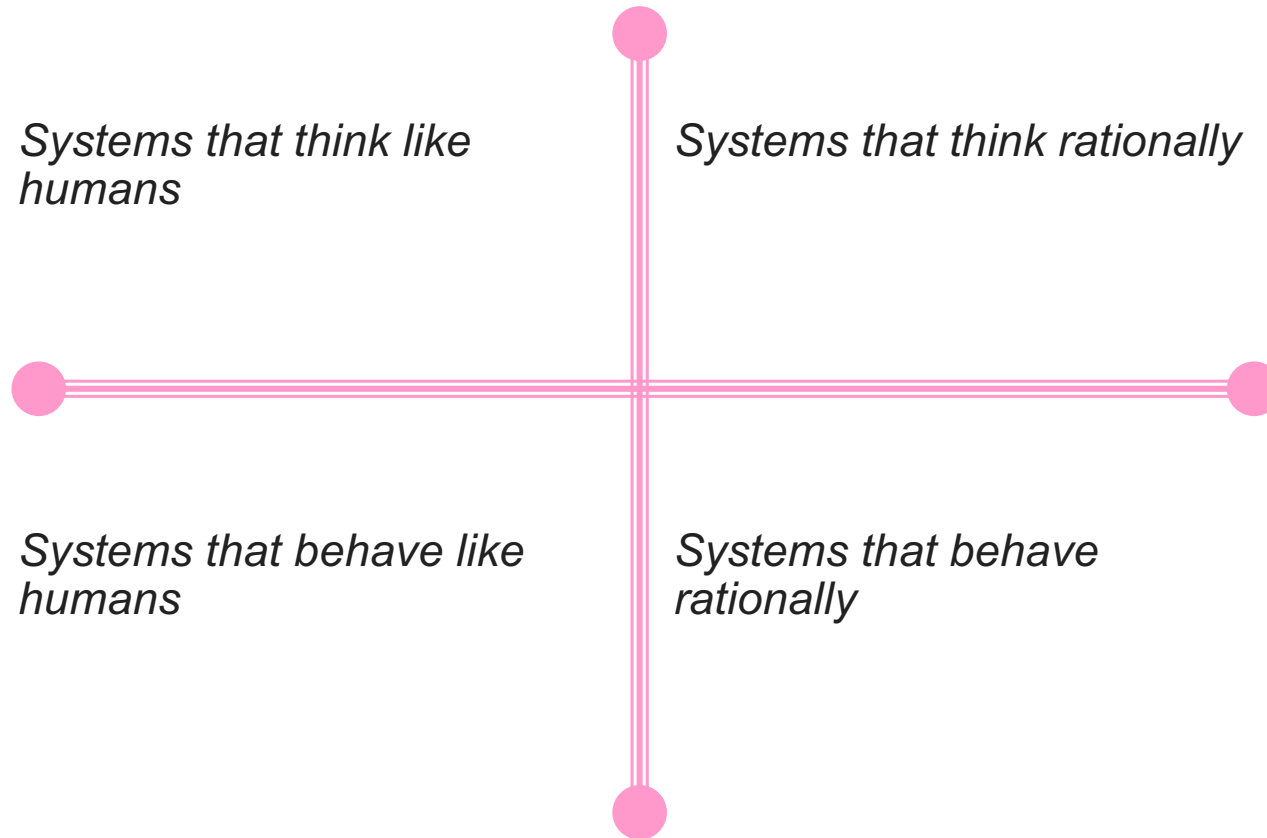


Bertrand Russell
(1872-1970)



Carnap, Rudolf
(1891-1970)

The AI we envisioned



The AI envisioned

Systems that think like humans

Systems that think rationally

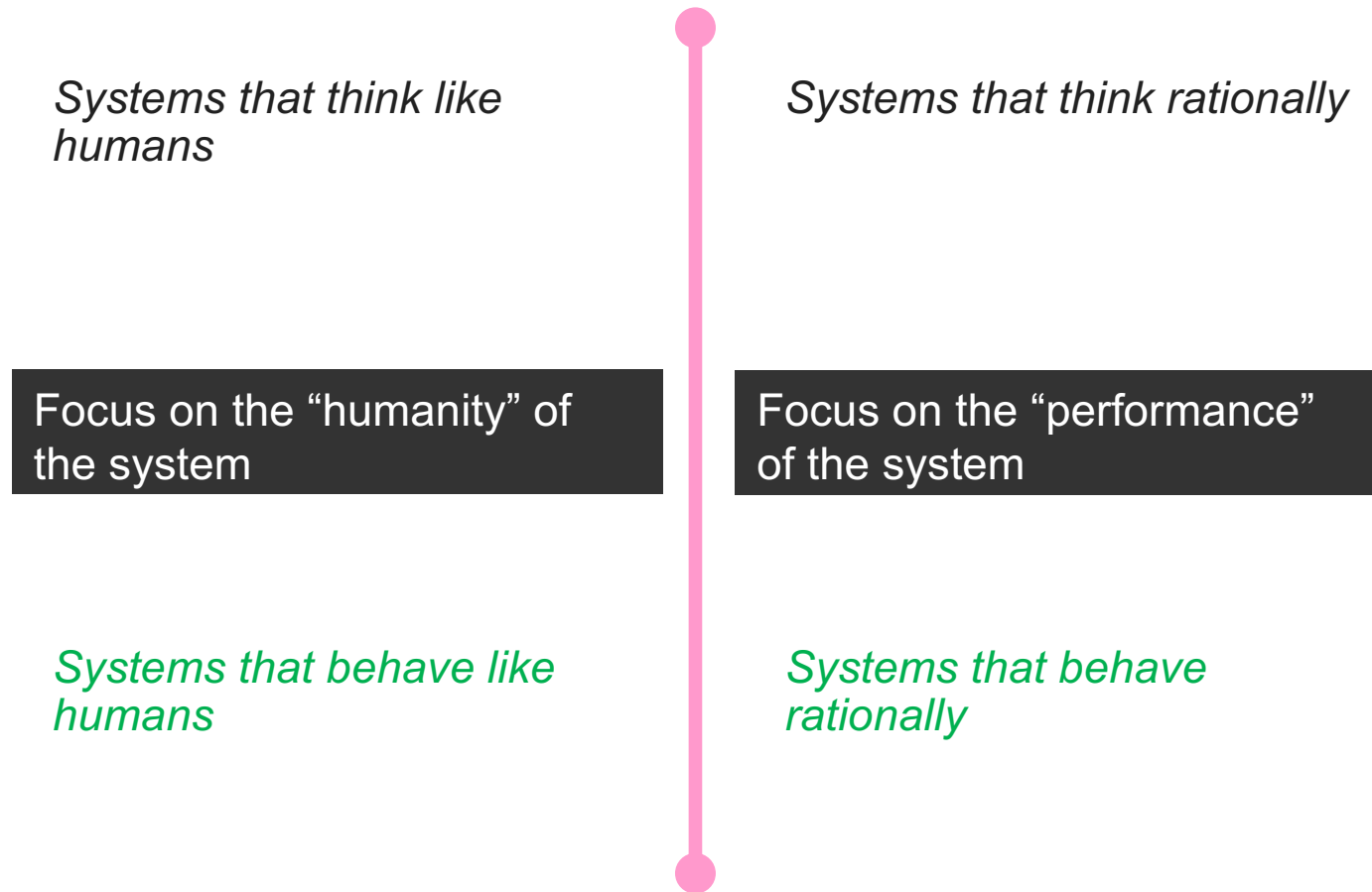
Focus on mental and reasoning processes

Focus on behavior

Systems that behave like humans

Systems that behave rationally

The AI envisioned



The AI envisioned

***Systems that
think like
humans***

Systems that think rationally

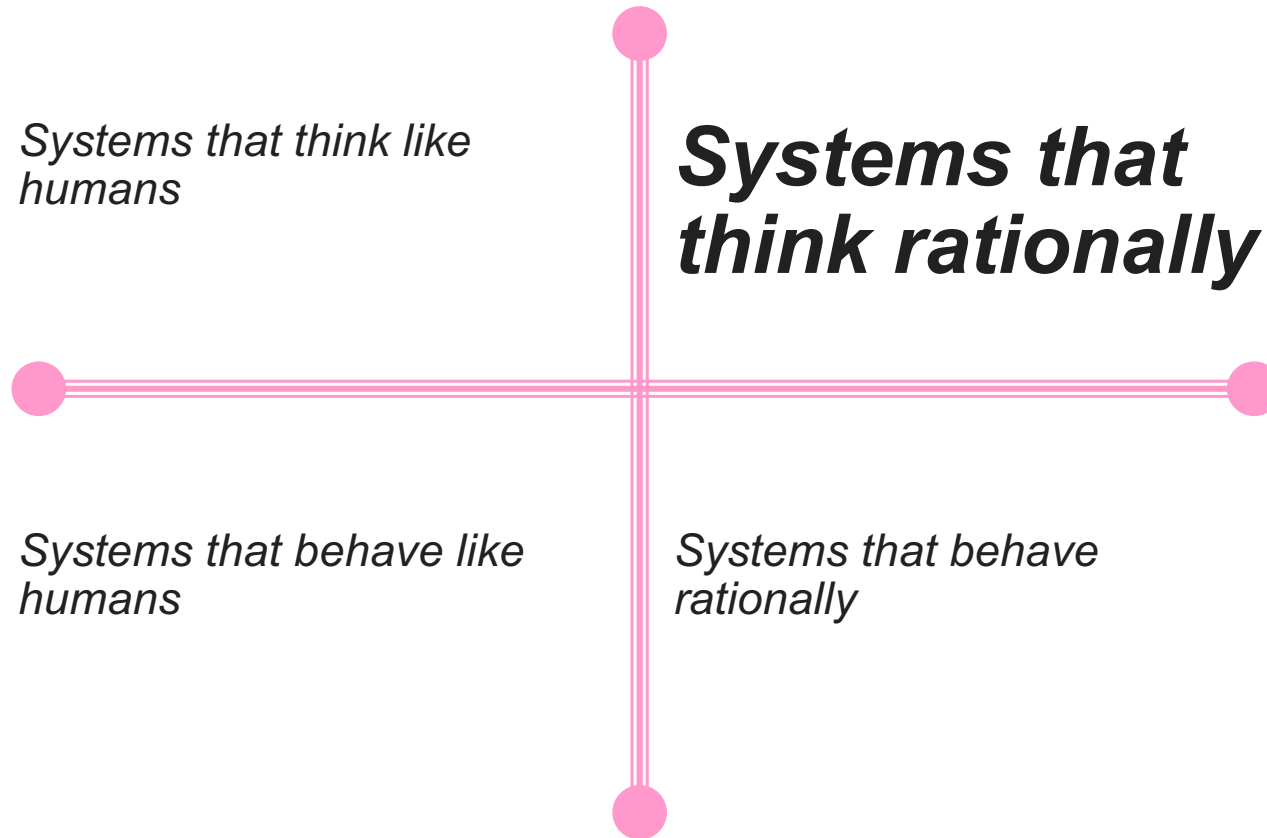
*Systems that behave like
humans*

*Systems that behave
rationally*

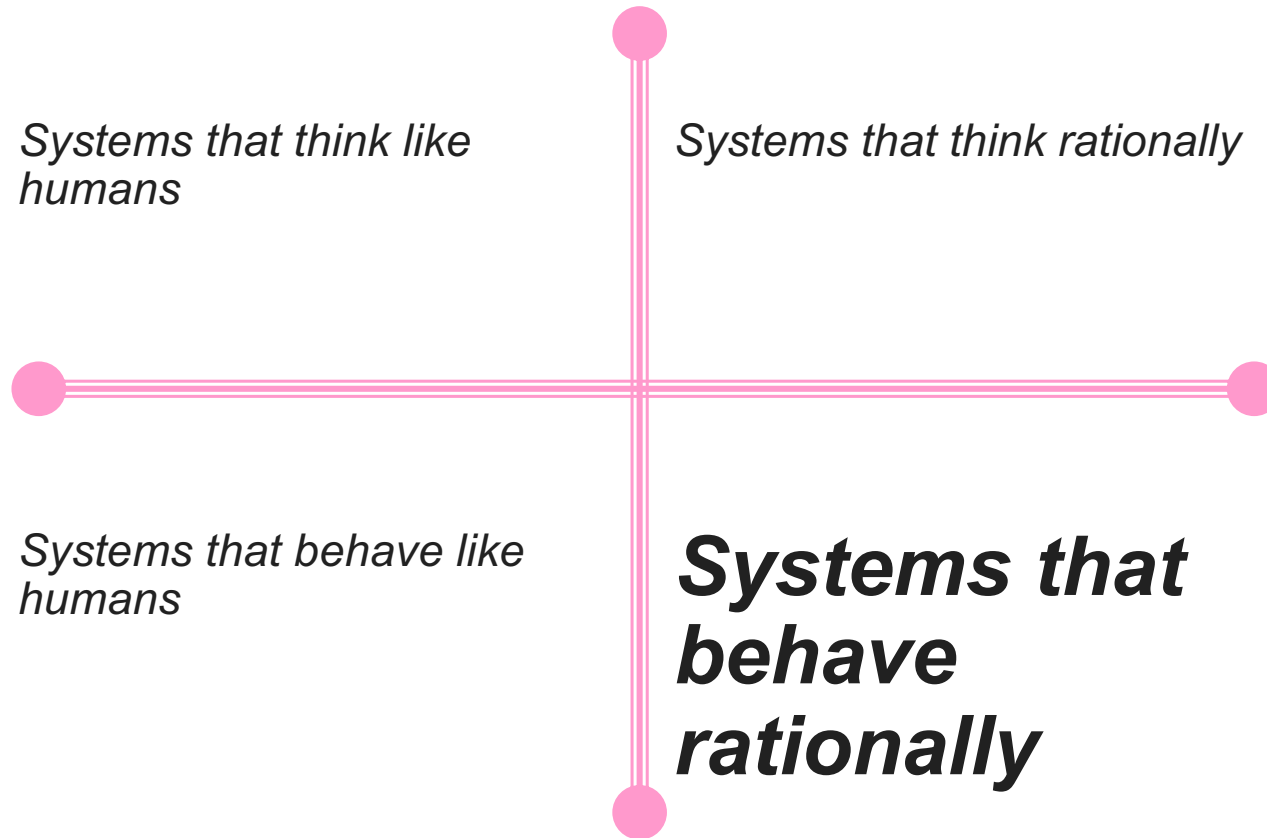
Thinking like humans

- This is the *cognitive* modeling approach.
- To achieve this modeling it is necessary to precisely understand the functioning of the *human mind*.
- That is, developing a theory of human thought that allows it to be accurately modeled.
- Modeling encompasses both human “hardware” and “software”, brain and mind.
- It is within the field of **Cognitive Science**.

The AI we envisioned



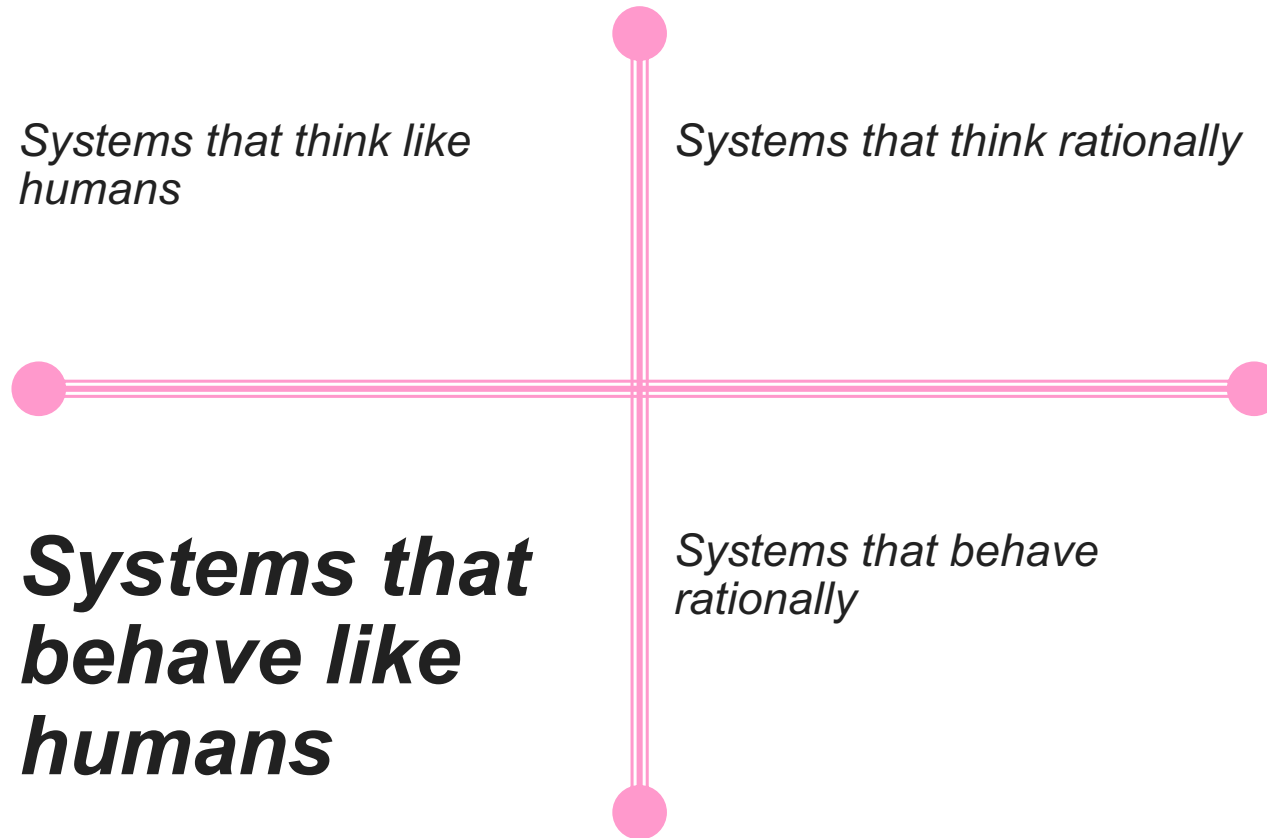
The AI we envisioned



Behaving rationally

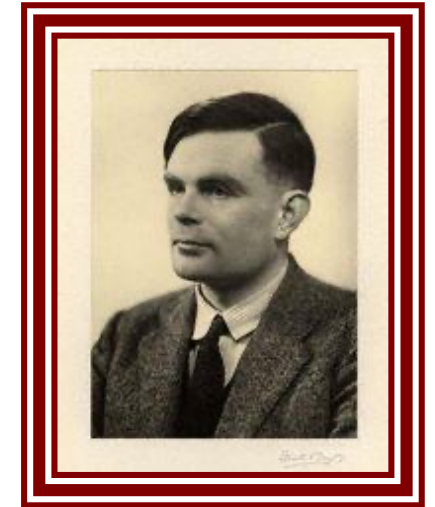
- Behaving rationally means acting in a way to achieve desired goals in the context of one's beliefs.
- Of course, one way to act rationally is to think rationally and act upon that conclusion.
- However, there may be cases where there is no provable correct action (because of lack of knowledge, lack of resources, etc), but an agent is expected to act anyway.
- There are also rational acts that do not come from rational inferences, for example reflex actions.

The AI we envisioned



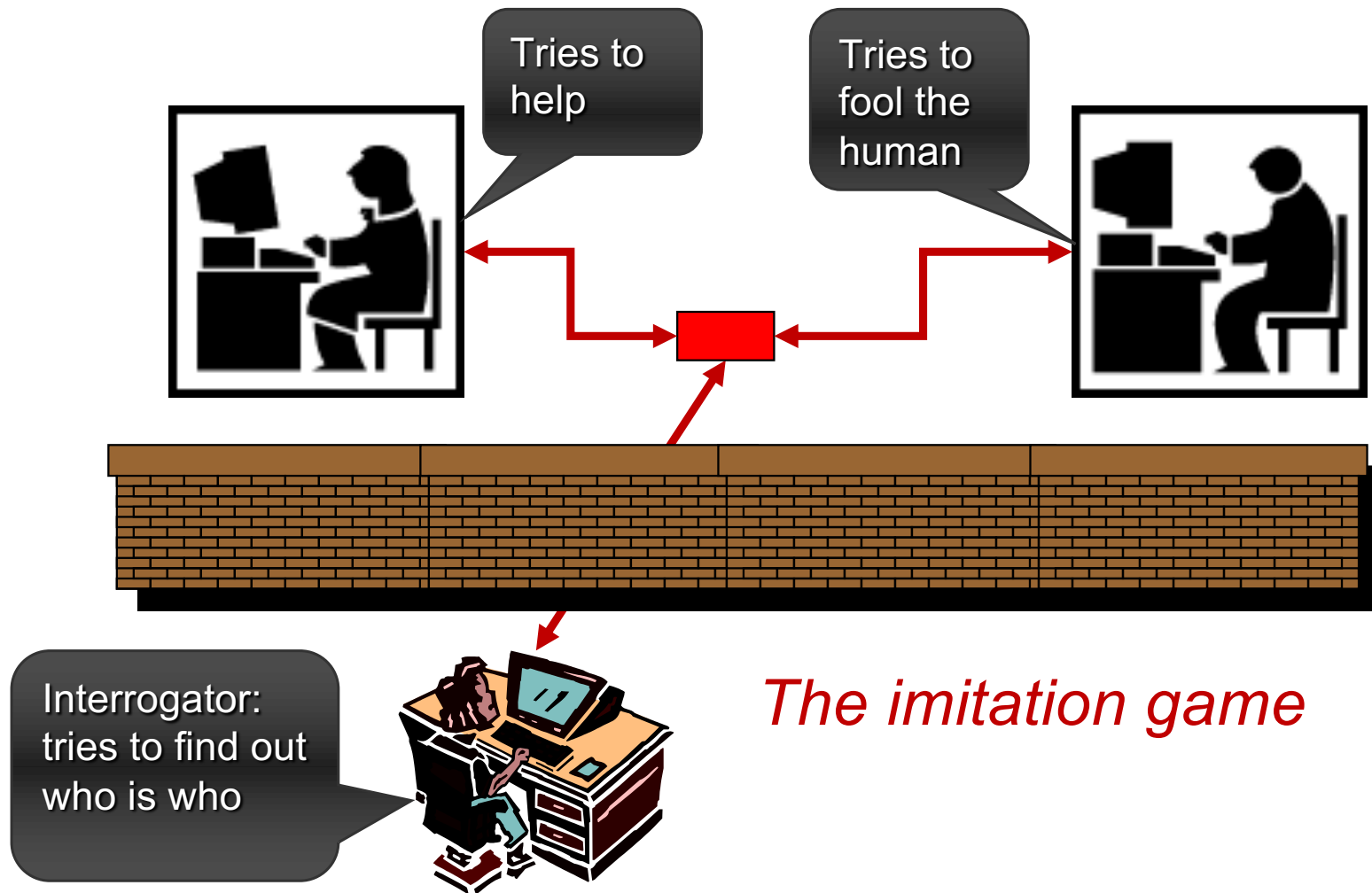
Behaving like humans

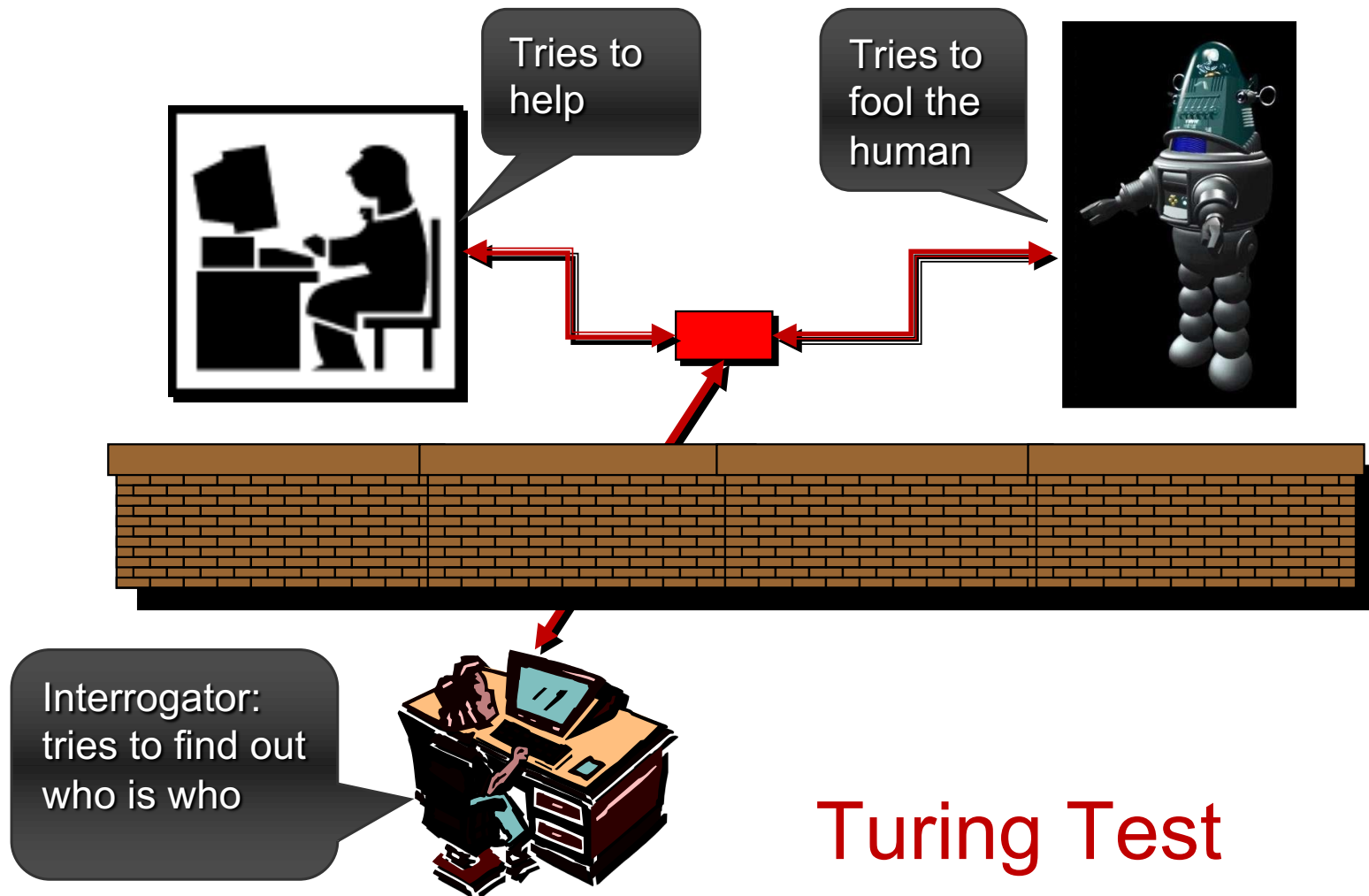
- The goal is to give an operational definition of intelligence.
- We define intelligence based on the class of **behavior** that the entity **possesses**.
- Alan Turing specified intelligent behavior as *“the ability to reach performance equivalent to human level in ALL cognitive tasks, sufficiently enough as to fool a human being”*, and suggested imitation game (Turing test).
- Turing test does not provide a definition of intelligence but a criteria of (average? Minimum?) human intelligence.



Alan Turing

“Can Machines Think?” Computing Machinery and Intelligence, Mind, Vol LIX, No 236 (1950) [7]





Turing Test

Definition vs Criteria

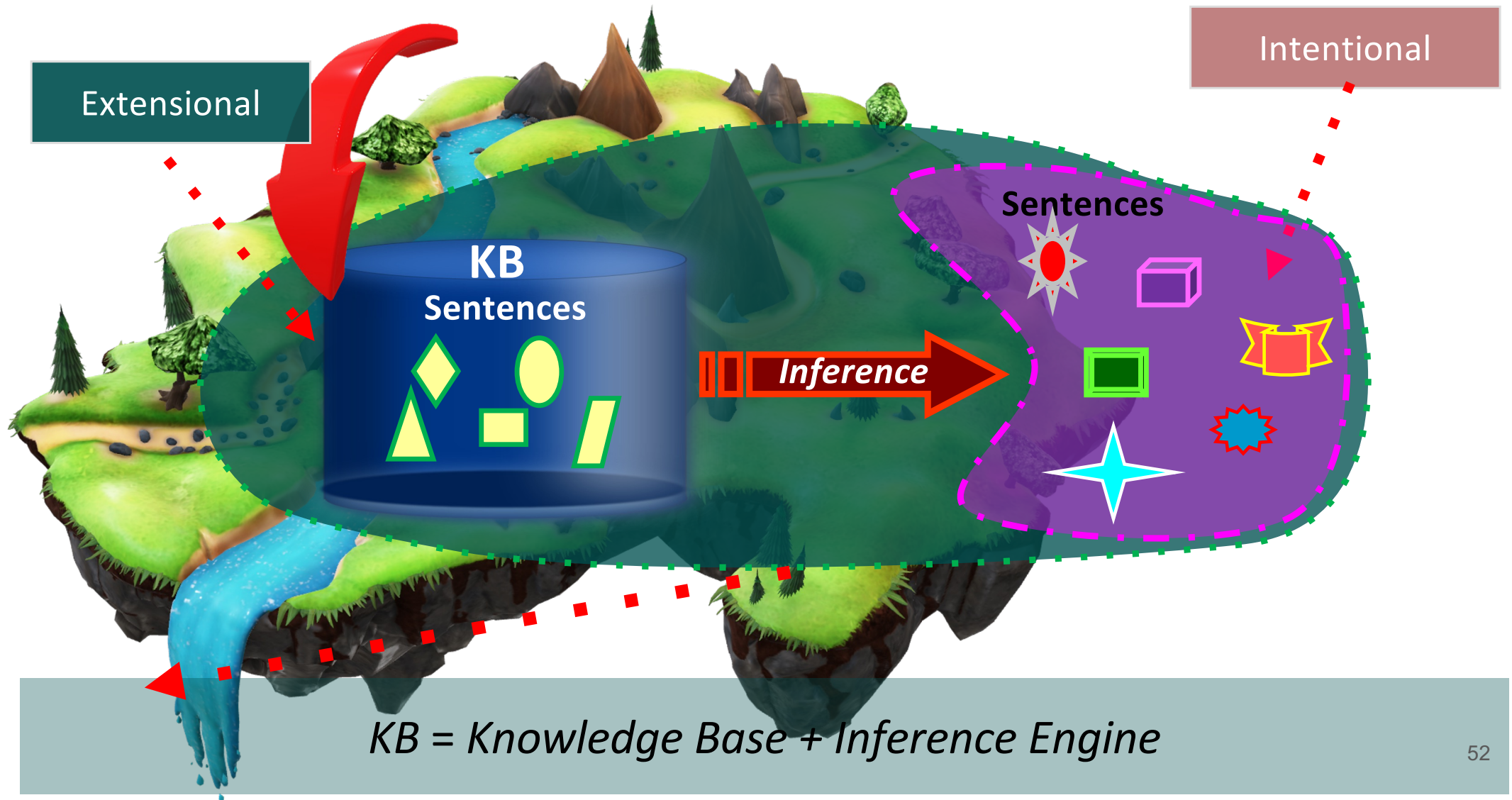
- A *definition* of intelligence would introduce an infallible way of separating intelligent things from the not intelligent ones.
- A *criteria* is a set of conditions or rules that allows to make a choice, which implies that based on that specific **criteria** it is possible to make decisions or put forward a value judgement. It is by definition *imprecise*.
- A criterion is composed of two parts: a parameter (the aspect that is measurable in the evaluation: the interaction) and the necessary standard regarding the parameter that is evaluated (the whole setting in which what we are measuring if how much the human perceives the machine as a human).

Knowledge Representation and Reasoning

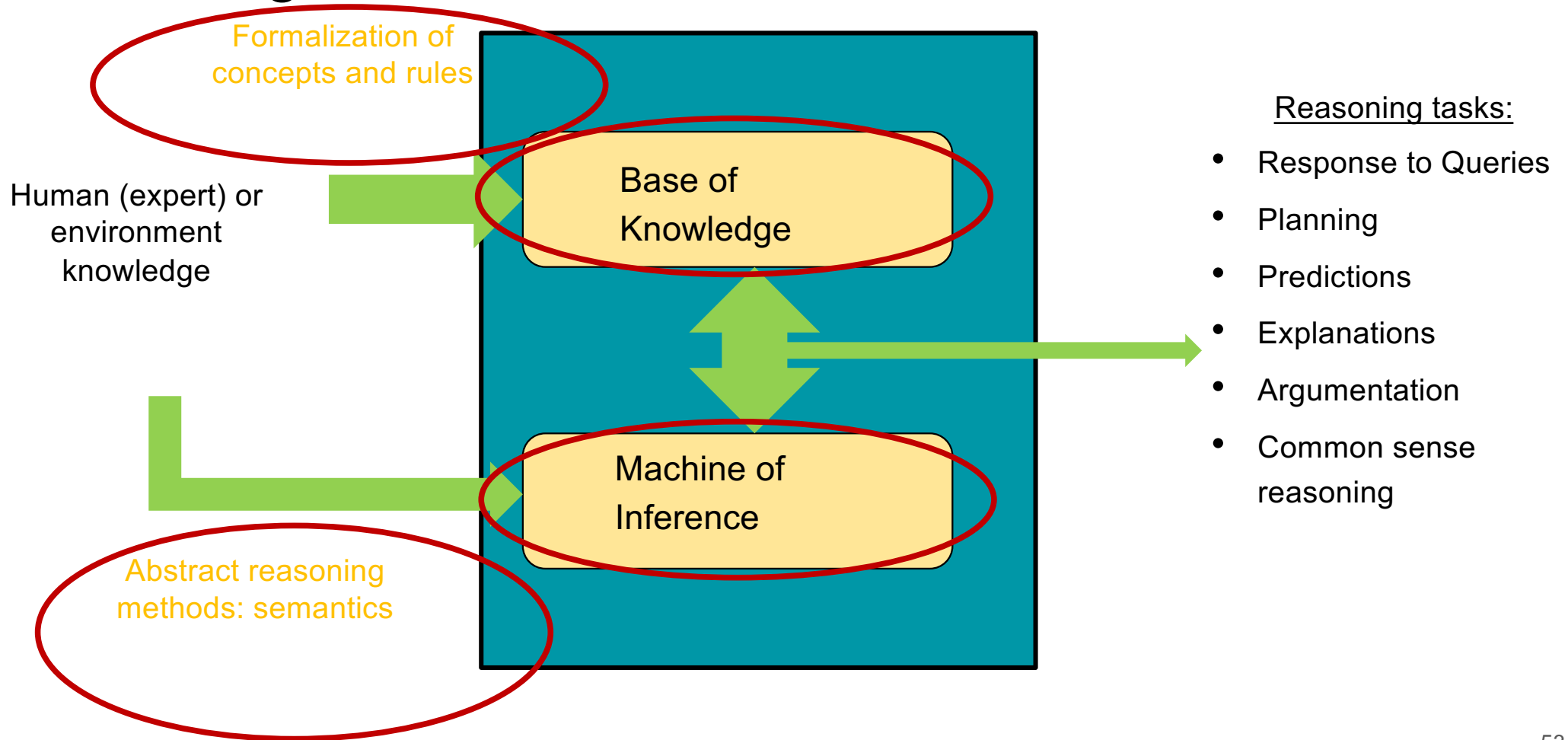
- Artificial intelligence systems that are supposed to perceive and act on the world need models of the part of the world they are supposed to work with.
- KR&R is the area of study that focuses on the problem of encoding human knowledge and reasoning (Automated Reasoning) into a symbolic language that can be processed by information systems.

Knowledge-based AI assumes two main tasks:

- Representation: To apply any type of reasoning, we need description of the scenario that we can manipulate *adequately*.
- Reasoning: Once we have the scenario represented, we can make inferences from it (for example, say things that are or are not explicit in the description).



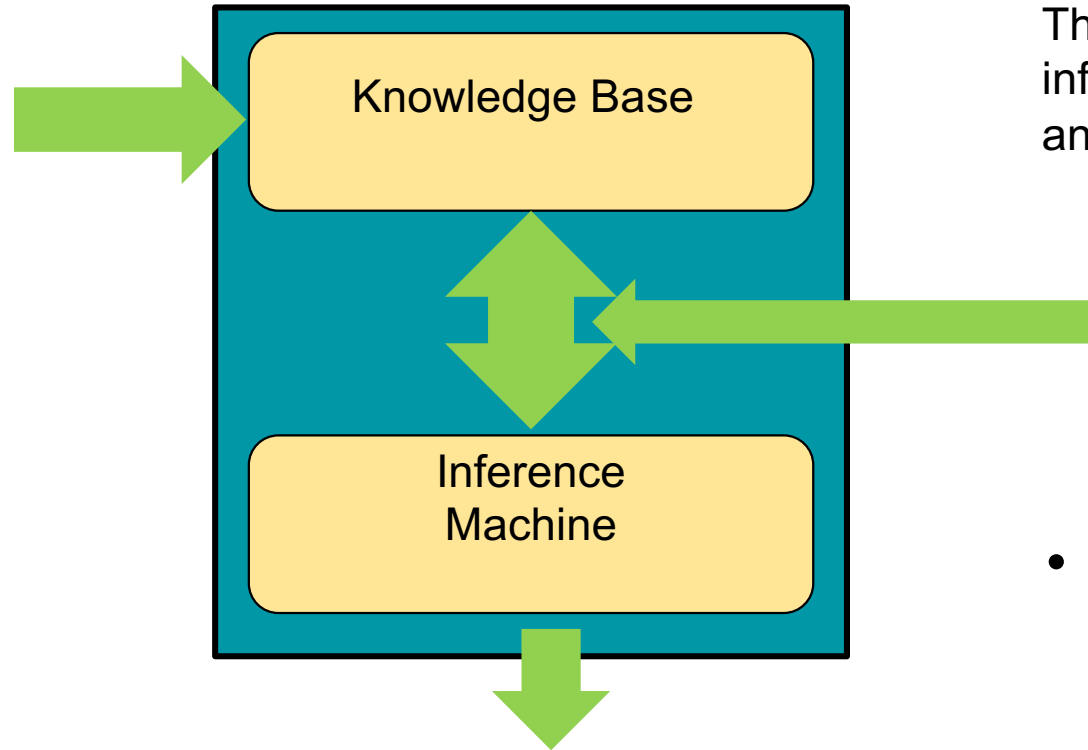
Knowledge-based AI



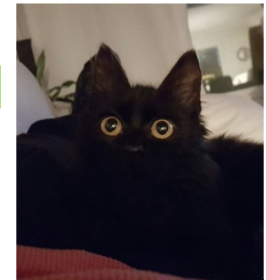
Knowledge-based AI: basic reasoning example

Human reports:

- “Dragons have wings”
- “Dragons fly”
- “Cats have hair”
- “Cats don’t fly”
-
- “Cats have tails”



The user provides some information about an animal named Homy.



- User asks: “Does Homy fly?”

“Homy doesn’t fly because he’s a cat.”

Which AI do we need?

The community understands that AI systems that support decision-making should:

- Involve higher-level reasoning.
- Form the “same” types of semantic representations and inferences that humans use and create.
- Have the ability to combine automatically acquired information and data with the domain-specific experience of users.

Which AI do we want?

- In addition, they must be legal, ethical, and robust
- Interpretable, predictable, (sometimes) explainable, etc.
- Aligned with human values.

Knowledge-based AI methods appear to have the tools to achieve these goals.

- Logic-based formalisms are especially interesting for modeling such skills in the form of normative reasoning.

Data vs. Knowledge: the original dichotomy

Knowledge-based AI:

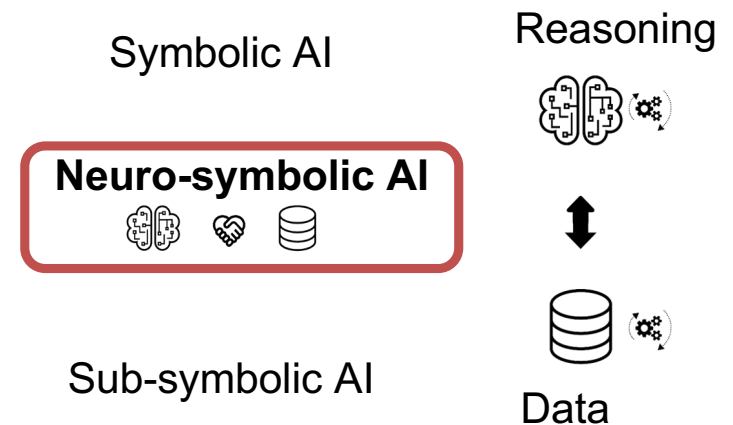
- Interpretable by design.
- It allows to represent knowledge of different kinds.
- It is not clear how to automatically get from low-level observations to high-level structured knowledge (knowledge is produced by humans).
- The hypothesis that humans use symbols to represent the knowledge needed to solve problems and we solve them by manipulating those symbols (we can formalize human reasoning) not necessarily applies to every type of cognitive task.
- They don't scale well...

Data vs. Knowledge: the original dichotomy

- Knowledge-based AI has proven to be most successful at intellectual tasks such as expert problem-solving: conceptual learning, planning, argumentation, decision-making, and general reasoning.
- Data-driven AI is most successful at tasks related to low-level cognitive tasks: pattern recognition such as images, voices, words, speaking skills, reading skills, robotic actions, language manipulation/generation.

Neuro-symbolic and Hybrid AI

- Aims to build a technical bridge between the two.
- Hybrid systems would:
 - be able to address the knowledge acquisition bottleneck faced by symbolic systems,
 - learn to perform logical reasoning tasks even in the presence of noise or uncertainty,
 - obtain self-explanatory data-driven models.
 - shed light on how natural (human) neural networks can perform mathematics, formal logic, and other activities that we, introspectively, see as symbolic in nature.



Neuro-symbolic reasoning

- “Neural Symbolic AI” (**NeSy AI**) is a family of approaches that combine connectionist (data-driven) and symbolic (knowledge-based) AI.
- NSR is the subset of NeSy that focuses on reasoning tasks.
- Kautz's Taxonomy for NeSy AI:
 1. Symbolic Neuro symbolic
 2. Symbolic [Neuro]
 3. Neuro | Symbolic
 4. Neuro:Symbolic®Neuro
 5. Neuro_{Symbolic}
 6. Neuro [Symbolic]

Two examples of hybrid AI

Automatic Generation of Hypotheses

- Using domain (eg., cybersecurity, health, finance) expert knowledge.
- Use information obtained from available data (machine learning, statistics analysis, etc.).
- Generation of hypothesis on potentially incomplete/unknown instances.
- Hypothesis validation from human analysts (*human in the loop*).

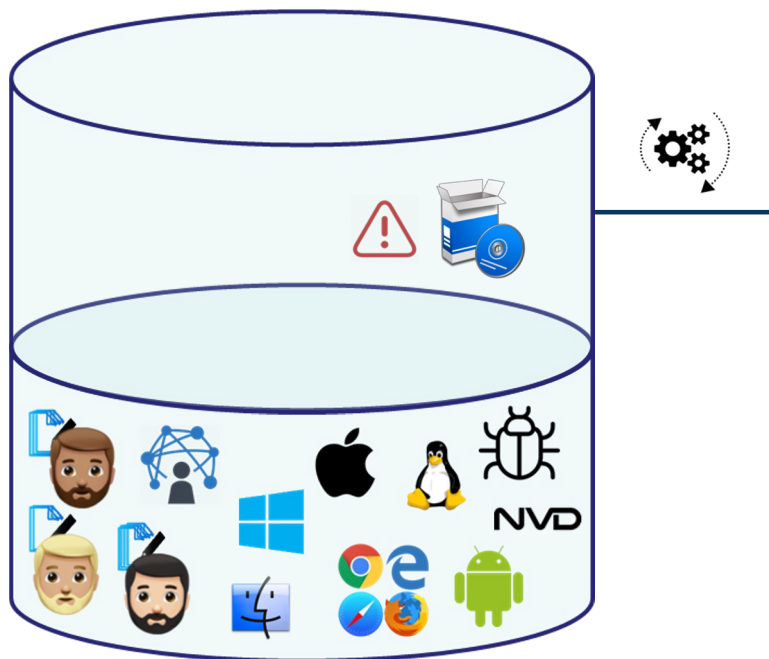
Automatic Generation of Hypotheses

Problem: *Given a system configuration, can we predict if/when the system may be at risk of suffering an attack?*

- Information regarding posts written by different actors, and their interactions on forums in online hacker communities such as the deepweb and darkweb:
CYR3CON (<https://cyr3con.ai>)
- Domain experts monitoring different sources of information.
- Incomplete/uncertain information (adversarial behavior).
- Many works focusing on data-driven approaches:
 - Hard to incorporate domain expert in the loop
 - don't provide explanations for analysts



Example: Threat Hunting with Hybrid AI



Using logical rules that feed off all this information, we can infer that, for example, there is a software (*unknown for now*) that is at risk of being attacked.

See: Paredes, Jose N., et al. "NetDER: an architecture for reasoning about malicious behavior." *Information Systems Frontiers* 23.1 (2021): 185-201.

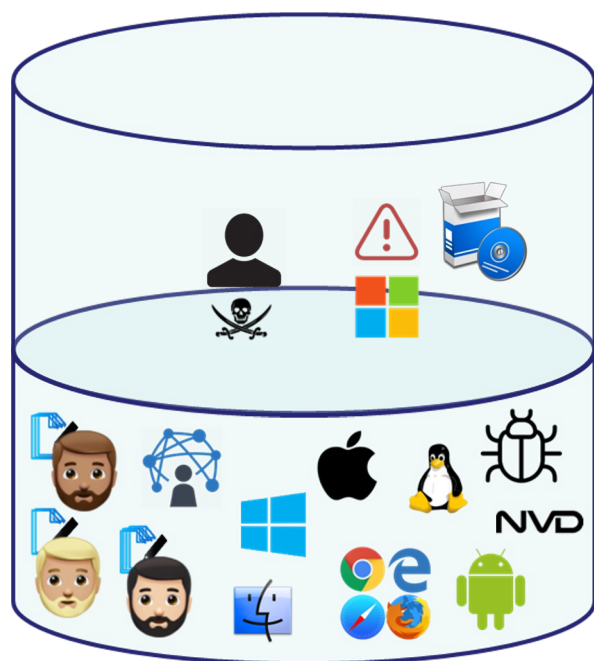
Paredes, Jose N., et al. "Detecting malicious behavior in social platforms via hybrid knowledge-and data-driven systems." FGCS 125 (2021): 232-246.

Paredes, Jose N., et al. "The HEIC application framework for implementing XAI-based socio-technical systems." *Online Social Networks and Media* 32 (2022): 100239

Teze, Juan Carlos L. et al. "Engineering user-centered explanations to query answers in ontology-driven socio-technical systems." *Semantic Web*, pre-press, 2024.



Example: *Threat Hunting with Hybrid AI*



Using logical rules that feed off all this information, we can infer that, for example, there is a *Microsoft software* (*unknown for now*) that is at risk of being attacked.

We could also hypothesize that *there is a user* (*unknown for now*) selling malware for this type of attack.

See: Paredes, Jose N., et al. "NetDER: an architecture for reasoning about malicious behavior." Information Systems Frontiers 23.1 (2021): 185-201.

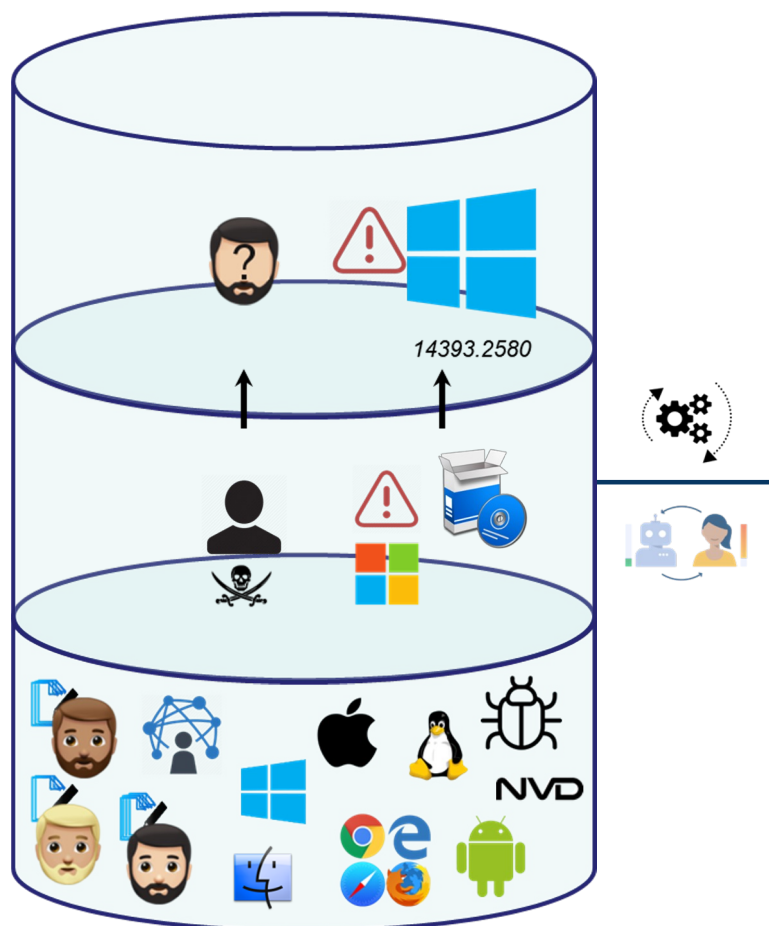
Paredes, Jose N., et al. "Detecting malicious behavior in social platforms via hybrid knowledge-and data-driven systems." FGCS 125 (2021): 232-246.

Paredes, Jose N., et al. "The HEIC application framework for implementing XAI-based socio-technical systems." Online Social Networks and Media 32 (2022): 100239

Teze, Juan Carlos L. et al. "Engineering user-centered explanations to query answers in ontology-driven socio-technical systems." Semantic Web, pre-press, 2024.



Opportunities



Based on this, and with the intervention of human analysts, a possible outcome is:

- The software at risk is *Windows 10* (build 14393.2580)
- *deduplication* hypothesis under uncertainty.

Analysts can then evaluate these inferences and possibly introduce new information to *revise the results* obtained.

See: Paredes, Jose N., et al. "NetDER: an architecture for reasoning about malicious behavior." Information Systems Frontiers 23.1 (2021): 185-201.

Paredes, Jose N., et al. "Detecting malicious behavior in social platforms via hybrid knowledge-and data-driven systems." FGCS 125 (2021): 232-246.

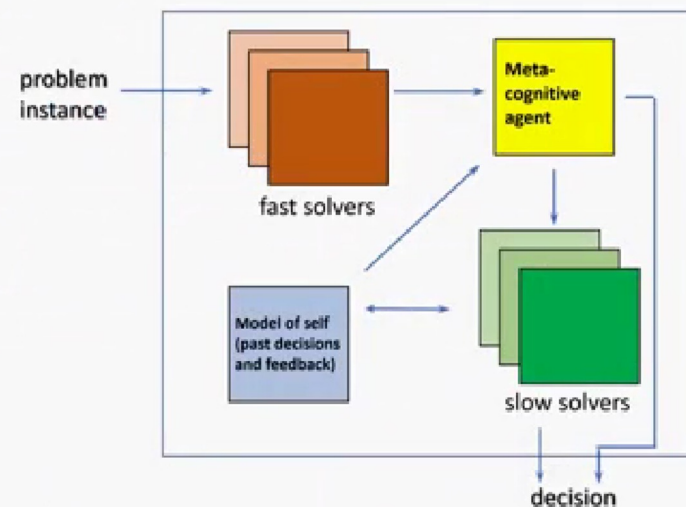
Paredes, Jose N., et al. "The HEIC application framework for implementing XAI-based socio-technical systems." Online Social Networks and Media 32 (2022): 100239

Teze, Juan Carlos L. et al. "Engineering user-centered explanations to query answers in ontology-driven socio-technical systems." Semantic Web, pre-press, 2024.

Source: <https://ibm.ent.box.com/s/75xnvwcw2r5ow0kmxg9fc8qsbulo9h8s>

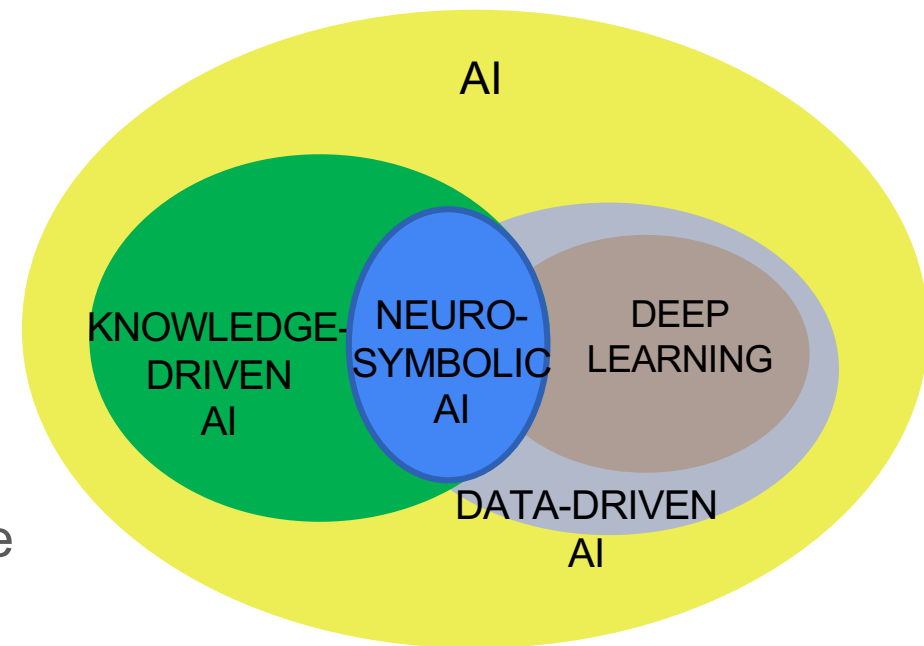
The SOFAI Architecture

- Multi-agent
- Fast solvers rely on past experience and act independently
- Slow solver reason on the problem and are prompted by the metacognitive agent
- Metacognitive (MC) agent decides whether to adopt fast solver's decision proposal or to deploy a slow solver
 - First, analysis of fast solver's proposal
 - If needed, comparison between expected reward and cost of the two kinds of solvers



Wrapping up

- The current narrative around AI make us focus and push into one specific direction, narrowing our future: huge data + huge models.
- A broader picture could help us develop a better technology developed within better frameworks of control.
- AI is not given, it is designed, developed, built by humans: we can build the AI that we want and more importantly the one that we need.



Thanks!
Questions?



References

1. <https://www.iiia.csic.es/barcelonadeclaration/>
2. Computational Intelligence D. Poole, A. Mackworth, R. Goebel.
3. Commonsense Reasoning, Erik T. Muller, Elsevier 2006
4. IBM Neuro-Symbolic AI Workshop 2023: <https://ibm.github.io/neuro-symbolic-ai/blog/nsai-wkshp-2023-blog/>
5. Thinking Fast and Slow in AI, G. Booch, F. Fabiano, L. Horesh, K. Kate, J. Lenchner, N. Linck, A. Loreggia, K. Murugesan, N. Mattei, , F. Rossi, B. Srivastava
6. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.
7. Artur S. d'Avila Garcez, Luís C. Lamb, Dov M. Gabbay (auth.). Neural-Symbolic Cognitive Reasoning.
8. Neuro-symbolic Reasoning: <https://www.digitaltrends.com/cool-tech/neuro-symbolic-ai-the-future/>
9. (PDF) *Evasion and causative attacks with adversarial deep learning*.
<https://www.researchgate.net/publication/321739488>.